PeerJ

Parcellating connectivity in spatial maps

Christopher Baldassano[1], Diane M. Beck[2] and Li Fei-Fei[1]

[1] Department of Computer Science, Stanford University, Stanford, CA, USA
[2] Beckman Institute and Department of Psychology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

ABSTRACT

A common goal in biological sciences is to model a complex web of connections using a small number of interacting units. We present a general approach for dividing up elements in a spatial map based on their connectivity properties, allowing for the discovery of local regions underlying large-scale connectivity matrices. Our method is specifically designed to respect spatial layout and identify locally-connected clusters, corresponding to plausible coherent units such as strings of adjacent DNA base pairs, subregions of the brain, animal communities, or geographic ecosystems. Instead of using approximate greedy clustering, our nonparametric Bayesian model infers a precise parcellation using collapsed Gibbs sampling. We utilize an infinite clustering prior that intrinsically incorporates spatial constraints, allowing the model to search directly in the space of spatially-coherent parcellations. After showing results on synthetic datasets, we apply our method to both functional and structural connectivity data from the human brain. We find that our parcellation is substantially more effective than previous approaches at summarizing the brain's connectivity structure using a small number of clusters, produces better generalization to individual subject data, and reveals functional parcels related to known retinotopic maps in visual cortex. Additionally, we demonstrate the generality of our method by applying the same model to human migration data within the United States. This analysis reveals that migration behavior is generally influenced by state borders, but also identifies regional communities which cut across state lines. Our parcellation approach has a wide range of potential applications in understanding the spatial structure of complex biological networks.

Subjects Computational Biology, Neuroscience, Computational Science
Keywords Parcellation, Clustering, Connectivity, fMRI, Brain, Tractography, Connectome, Probabilistic model, Migration, Spatial maps

Submitted 18 December 2014
Accepted 1 February 2015
Published 19 February 2015

Corresponding author
Christopher Baldassano, chrisb33@cs.stanford.edu

Academic editor
Yong Wang

Additional Information and Declarations can be found on page 19

DOI **10.7717/peerj.784**

INTRODUCTION

When studying biological systems at any scale, scientists are often interested not only in the properties of individual molecules, cells, or organisms, but also in the web of *connections* between these units. The rise of massive biological datasets has enabled us to measure these second-order interactions more accurately, in domains ranging from protein–protein interactions, to neural networks, to ecosystem food webs. We can often gain insight into the overall structure of a connectivity graph by grouping elements into clusters based on their connectivity properties. Many types of biological networks have been modeled in terms of interactions between a relatively small set of "modules" (*Barabási & Oltvai,*

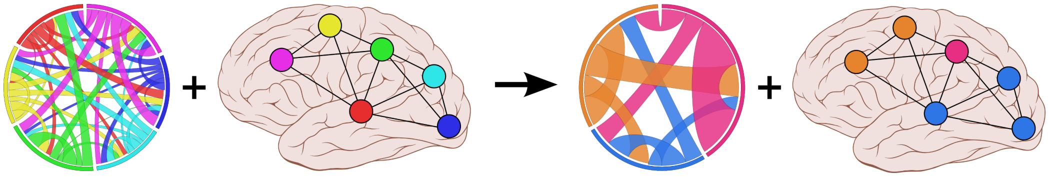

Figure 1 Parcellating connectivity in spatial maps Given a set of elements arranged on a spatial map (such as points within the human cortex) as well as the connectivity between each pair of elements, our method finds the best parcellation of the spatial map into connected clusters of elements that all have similar connectivity properties. Brain image by Patrick J. Lynch, licensed under CC BY 2.5.

2004; Hartwell et al., 1999), including protein–protein interactions (Rives & Galitski, 2003), metabolic networks (Ravasz et al., 2002), bacterial co-occurrence (Freilich et al., 2010), pollination networks (Olesen et al., 2007), and food webs (Krause, Frank & Mason, 2003). In fact, it has been proposed that modularity may be a necessary property for any network that must adapt and evolve over time, since it allows for reconfiguration (Alon, 2003; Hartwell et al., 1999). There are a large number of methods for clustering connectivity data, such as k-means (Kim et al., 2010; Golland et al., 2008; Lee et al., 2012), Gaussian mixture modeling (Golland, Golland & Malach, 2007), hierarchical clustering (Mumford et al., 2010; Cordes et al., 2002; Gorbach et al., 2011), normalized cut (Van den Heuvel, Mandl & Hulshoff Pol, 2008), infinite relational modeling (Morup et al., 2010), force-directed graph layout (Crippa et al., 2011), weighted stochastic block modeling (Aicher, Jacobs & Clauset, 2014), and self-organized mapping (Mishra et al., 2014; Wiggins et al., 2011).

The vast majority of these methods, however, ignore the fact that biological networks almost always have some underlying spatial structure. As described by Legendre and Fortin: "In nature, living beings are distributed neither uniformly nor at random. Rather, they are aggregated in patches, or they form gradients or other kinds of spatial structures…the spatio-temporal structuring of the physical environment induces a similar organization of living beings and of biological processes, spatially as well as temporally" (Legendre & Fortin, 1989). In many biological datasets, we therefore wish to constrain possible clustering solutions to consist of *spatially-contiguous parcels*. For example, when dividing a DNA sequence into protein-coding genes, we should enforce that the genes are contiguous sequences of base pairs. Similarly, if we want to identify brain regions that could correspond to local cortical modules, we need each discovered cluster to be a spatially-contiguous region. Without spatial information, the discovered clusters may be difficult to interpret; for example, clustering functional brain connectivity data without spatial information yields spatially-distributed clusters that confound local modularity and long-distance interactions (Lee et al., 2012).

The problem is thus to a parcellate a spatial map into local, contiguous modules such that all elements in a module have the same connectivity properties (Fig. 1). In this paper we present the first general solution to this problem, introducing a new generative probabilistic model to parcellate a spatial map into local regions with connectivity properties that are as uniform as possible. Scientific insights can be gained from both the clusterings themselves (which identify the local spatial sources of the interaction matrix) as well as the connections between the parcels, which summarize the original complex connectivity matrix. Our method yields better results than other approaches such as greedy clustering, and can help to determine the correct number of parcels in a data-driven way.

One of the most challenging spatial parcellation problems is in the domain of neuroscience. Modern human neuroimaging methods can estimate billions of connections between different locations in the brain, with complex spatial structures that are highly nonuniform in size and shape. Correctly identifying the detailed boundaries between brain regions is critical for understanding distributed neural processing, since even small inaccuracies in parcellation can yield major errors in estimating network structure (*Smith et al., 2011*).

Obtaining a brain parcellation with spatially coherent clusters has been difficult, since it is unclear how to extend standard clustering methods to include the constraint that only adjacent elements should be clustered together. Biasing the connectivity matrix to encourage local solutions can produce local parcels in some situations (*Cheng & Fan, 2014*; *Tomassini et al., 2007*), or distributed clusters can be split into their connected components after clustering (*Abraham et al., 2013*), but these approximations will not necessarily find the best parcellation of the original connectivity matrix. It is also possible to add a Markov Random Field prior (such as the Ising model) onto a clustering model to encourage connected parcels (*Jbabdi, Woolrich & Behrens, 2009*; *Ryali et al., 2013*), but in practice this does not guarantee that clusters will be spatially connected (*Honnorat et al., 2014*).

Currently, finding spatially-connected parcels is often accomplished using agglomerative clustering (*Thirion et al., 2014*; *Heller et al., 2006*; *Blumensath et al., 2013*; *Moreno-Dominguez, Anwander & Knosche, 2014*), which iteratively merges neighboring elements based on similarity in their connectivity maps. There are a number of disadvantages to this approach; most critically, the solution is only a greedy approximation (only a single pass over the data is made, and merged elements are never unmerged), which as will be shown below can lead to poor parcellations when there is a high level of noise. Edge detection methods (*Cohen et al., 2008*; *Wig, Laumann & Petersen, 2014*; *Gordon et al., 2014*) define cluster boundaries based on sharp changes in connectivity properties, which are also sensitive to localized patches of noisy data. Spectral approaches such as normalized cut (*Craddock et al., 2012*) attempt to divide the spatial map into clusters by maximizing within-cluster similarity and between-cluster dissimilarity, but this approach has a strong bias to choose clusters that all have similar sizes (*Blumensath et al., 2013*). It is also possible to incorporate a star-convexity prior into an MRF to efficiently identify connected parcels (*Honnorat et al., 2014*). This approach, however, constrains clusters to be convex (in connectivity space); as will be shown below, our method finds structures in real datasets violating this assumption, such as nested regions in functional brain connectivity data. All of these methods require explicitly setting the specific number of desired clusters, and are optimizing a somewhat simpler objective function; they seek to maximize the similarity between the one-dimensional rows or columns of the connectivity matrix, while our method takes into account reordering of the both the rows and columns to make the between-parcel 2D connectivity matrix as simple as possible.

Our model is highly robust to noise, has no constraints on the potential sizes and shapes of brain regions, and makes many passes over the data to precisely identify region boundaries. We validate that our method outperforms previous approaches on synthetic datasets, and then show that we can more efficiently summarize both functional and structural brain connectivity data. Our parcellation of human cortex generalizes more effectively across subjects, and reveals new structure in the functional connectivity properties of visual cortex.

To demonstrate the wide applicability of our method, we apply the same model to find spatial patterns in human migration patterns within the United States. Despite the fact that this is an entirely different type of data at a different spatial scale, we are able to find new insights into how state borders shape migratory behavior. Our results on these diverse datasets suggest that our analysis could have a wide range of potential applications in understanding biological networks. It is also important to note that the "spatial adjacency" constraint of our method could also be used for other, nonspatial notions of adjacency; for example, clustering an organism's life into contiguous temporal segments based on its changing social interactions.

MATERIALS AND METHODS

Probabilistic model

Intuitively, we wish to find a parcellation **z** which identifies local regions, such that all elements in a region have the same connectivity "fingerprint." Specifically, for any two parcels m and n, all pairwise connectivities between an element in parcel m and an element in parcel n should have a similar value. Our method uses the full distribution of all pairwise connectivities between two parcels, and finds a clustering for which this distribution is highly peaked. This makes our method much more robust than approaches which greedily merge similar clusters (*Thirion et al., 2014*; *Blumensath et al., 2013*) or define parcel edges where neighboring voxels differ (*Thirion et al., 2006*; *Wig, Laumann & Petersen, 2014*; *Gordon et al., 2014*). The goal of identifying modules with similar connectivity properties is conceptually similar to weighted stochastic block models (*Aicher, Jacobs & Clauset, 2014*), but it is unclear how these models could be extended to incorporate the spatial-connectivity constraint.

We would like to learn the number of regions automatically from data, and additionally impose the requirement that all regions must be spatially-connected. We can accomplish both goals more efficiently in a single framework, by using an infinite clustering prior on our parcellation **z** which simultaneously constrains regions to be spatially coherent and does not limit the number of possible clusters. Specifically, since the mere existence of a element (even with unknown connectivity properties) changes the spatial connectivity and thus affects the most likely clustering, we must employ a nonparametric prior which is *not marginally invariant*. Other Bayesian nonparametric models allow for spatial dependencies between datapoints, but the only class of CRPs which is not marginally invariant is the distance-dependent Chinese Restaurant Process (dd-CRP) (*Blei & Frazier, 2011*). Instead of directly sampling a label for each element, the dd-CRP prior assigns each element i a link to a neighboring element c_i. The actual parcel labels $\mathbf{z}(\mathbf{c})$ are then defined implicitly as the undirected connected components of the link graph. Intuitively, this allows for changes in the labels of many elements when a single connection c_i is modified, since it may break apart or merge together two large connected sets of elements. Additionally, this construction allows the model to search freely in the space of parcel links $\mathbf{c}$, since every possible setting of the parcel links corresponds to a parcellation satisfying the spatial-coherence constraint.

Given a parcellation, we must then specify a generative model for the data matrix $\mathbf{D}$. Analogous to the approach taken in stochastic block modeling (*Aicher, Jacobs & Clauset, 2014*), we model the connectivity between each pair of parcels as a separate distribution with latent parameters. To allow efficient collapsed sampling (see below), we utilize a Normal distribution for each set of connectivities between parcels, and the conjugate prior for the latent parameters.

Mathematically, our generative clustering model is:

$$\mathbf{c} \sim \text{dd-CRP}(\alpha, f)$$
$$A_{mn}, \sigma^2_{mn} \sim \text{Normal-Inverse-}\chi^2(\mu_0, \kappa_0, \sigma^2_0, \nu_0)$$
$$D_{ij} \sim \text{Normal}(A_{\mathbf{z}(\mathbf{c})_i \mathbf{z}(\mathbf{c})_j}, \sigma^2_{\mathbf{z}(\mathbf{c})_i \mathbf{z}(\mathbf{c})_j}).$$

For N elements and K parcels: $\mathbf{c}$ is a vector of length N which defines the cluster links for all elements (producing a region labeling vector $\mathbf{z}(\mathbf{c})$ of length N, taking values from 1 to K); α and f are the scalar hyperparameter and $N \times N$ distance function defining the dd-CRP; $\mathbf{A}$ and $\boldsymbol{\sigma}^2$ are the $K \times K$ connectivity strength and variance between regions; μ_0 and κ_0 are the scalar prior mean and precision for the connectivity strength; σ^2_0 and ν_0 are the scalar prior mean and precision for the connectivity variance; and $\mathbf{D}$ is the $N \times N$ observed connectivity between individual elements.

The probability of choosing a particular c_i in the dd-CRP is defined by a distance function f; we use $f_{ij} = 1$ if i and j are neighbors, and 0 otherwise, which guarantees that all clusters will be spatially connected. A hyperparameter α controls the probability that a voxel will choose to link to itself. Note that, due to our choice of distance function f, a random partition drawn from the dd-CRP can have many clusters even for $\alpha = 0$, since elements are only locally connected.

The connectivity strength A_{mn} and variance σ^2_{mn} between each pair of clusters m and n is given by a Normal-Inverse-χ^2 (NIχ^2) distribution, and the connectivity D_{ij} between every element i in one region and j in the other is sampled based on this strength and variance. The conjugacy of the Normal-Inverse-χ^2 and Normal distributions allows us to collapse over A_{mn} and σ^2_{mn} and sample only the clustering variables c_i. Empirically, we find that the only critical hyperparameter is the expected variance σ^2_0, with lower values encouraging parcels to be smaller (we set $\alpha = 10, \mu_0 = 0, \kappa_0 = 0.0001, \nu_0 = 1$ for all experiments).

To allow the comparison of hyperparameter values between problems with the same number of elements (e.g., the functional and structural datasets), we normalize the input matrix D to have zero mean and unit variance. We then initialize the model using the Ward clustering (see below) with the most likely number of clusters under our model, and setting the links $\mathbf{c}$ to form a random spanning tree within each cluster.

In summary, we have introduced a novel connectivity clustering model which (a) uses the full distribution of connectivity properties to define the parcellation likelihood, and (b) employs an infinite clustering model which automatically chooses the number of parcels and enforces that parcels be spatially-connected.

Derivation of Gibbs sampling equations

To infer a maximum a posteriori (MAP) parcellation $\mathbf{z}$ based on the dd-CRP prior, we perform collapsed Gibbs sampling on the element links $\mathbf{c}$. A link c_i for element i is drawn from

$$p(c_i^{(new)}|\mathbf{c_{-i}},D) \propto p(c_i^{(new)})p(D|\mathbf{z}(\mathbf{c_{-i}} \cup c_i^{(new)})) = p(c_i^{(new)})p(D|\mathbf{z}^{(new)})$$
$$\propto \begin{Bmatrix} \alpha & \text{if } c_i^{(new)} = i \\ 1 & \text{else} \end{Bmatrix} \prod_{k_1,k_2=1}^{|\mathbf{z}^{(new)}|} p(D_{z_{k_1}^{(new)}, z_{k_2}^{(new)}}). \tag{1}$$

To compare the likelihood term for different choices of $c_i^{(new)}$, we first remove the current link c_i, giving the induced partition $\mathbf{z}(\mathbf{c_{-i}})$ (which may split a region). If we resample c_i to a self-loop or to a neighbor j that does not join two regions, the likelihood term is based on the partition $\mathbf{z}(\mathbf{c_{-i}}) = \mathbf{z}$. Alternatively, c_i can be resampled to a neighbor j such that two regions K' and K'' in $\mathbf{z}(\mathbf{c_{-i}})$ are merged into one region K in $\mathbf{z}(\mathbf{c_{-i}} \cup c_i^{(new)}) = \hat{\mathbf{z}}$. Numbering the regions so that $z_i \in \{1 \cdots (K-1), K', K''\}$ and $\hat{z}_i \in \{1 \cdots (K-1), K\}$ gives

$$\frac{p(D|\hat{\mathbf{z}})}{p(D|\mathbf{z})} = \frac{\prod_{k=1}^{K} p(D_{\hat{z}_k,\hat{z}_K}) \prod_{k=1}^{K-1} p(D_{\hat{z}_K,\hat{z}_k})}{\prod_{k=1}^{K'} p(D_{z_k,z_{K'}}) \prod_{k=1}^{K''} p(D_{z_k,z_{K''}}) \prod_{k=1}^{K-1} p(D_{z_{K'},z_k}) \prod_{k=1}^{K'} p(D_{z_{K''},z_k})}. \tag{2}$$

Each term $p(D_{z_m,z_n})$ is a marginal likelihood of the NIχ^2 distribution, which can be computed in closed form (*Murphy, 2007*):

$$p(D_{z_m,z_n}) = \frac{\Gamma(\nu_{mn}/2)}{\Gamma(\nu_0/2)} \left(\frac{\kappa_0}{\kappa_{mn}}\right)^{\frac{1}{2}} \frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_{mn}\sigma_{mn}^2)^{\nu_{mn}/2}} (\pi)^{-n/2}$$

$$L = |z_m||z_n| \qquad \kappa_{mn} = \kappa_0 + L;\ \nu_{mn} = \nu_0 + L \qquad \mu_{mn} = \frac{\kappa_0\mu_0 + L\bar{d}}{\kappa_{mn}}$$

$$\bar{d} = \frac{1}{L} \sum_{\substack{i \in z_m \\ j \in z_n}} D_{ij} \qquad s = \sum_{\substack{i \in z_m \\ j \in z_n}} (D_{ij} - \bar{d})^2 \qquad \sigma_{mn}^2 = \frac{1}{\nu_{mn}} \left(\nu_0\sigma_0^2 + s + \frac{L\kappa_0}{\kappa_0 + L}(\mu_0 - \bar{d})^2\right).$$

Intuitively, Eq. (2) computes the probability of merging or splitting two regions at each step based on whether the connectivities between these regions' elements and the rest of the regions are better fit by one distribution or two.

In practice, the time-consuming portion of each sampling iteration is computing the sum of squared deviations s. This can be made more efficient by computing the s values for the merged $\hat{\mathbf{z}}$ in closed form. Given that the connectivities $D_{K'} = \{D_{iK'}\}_{i \in k}$ between parcel k

and K' have sum of squares deviations $s_{K'}$ and mean $\bar{d}_{K'}$, and similarly for K'', then the sum of squares s_K for the connectivities between parcel k and the merged parcel K (merging K' and K'') is:

$$\begin{aligned}
s_K &= \sum_{d\in D_{K'}\cup D_{K''}} (d-\bar{d})^2 \\
&= \left(\sum_{d\in D_{K'}\cup D_{K''}} d^2\right) - (|D_{K'}|+|D_{K''}|)\cdot\left(\frac{|D_{K'}|\cdot\bar{d}_{K'}+|D_{K''}|\cdot\bar{d}_{K''}}{|D_{K'}|+|D_{K''}|}\right)^2 \\
&= \left(\sum_{d\in D_{K'}\cup D_{K''}} d^2\right) - \frac{|D_{K'}|^2}{|D_{K'}|+|D_{K''}|}\bar{d}^2_{K'} - \frac{|D_{K''}|^2}{|D_{K'}|+|D_{K''}|}\bar{d}^2_{K''} - 2\frac{|D_{K'}||D_{K''}|}{|D_{K'}|+|D_{K''}|}\bar{d}_{K'}\bar{d}_{K''} \\
&= \left(\sum_{d\in D_{K'}} d^2 - |D_{K'}|\bar{d}^2_{K'}\right) + \left(\sum_{d\in D_{K''}} d^2 - |D_{K''}|\bar{d}^2_{K''}\right) \\
&\quad + \frac{|D_{K'}||D_{K''}|}{|D_{K'}|+|D_{K''}|}(\bar{d}^2_{K'}+\bar{d}^2_{K''}-2\bar{d}_{K'}\bar{d}_{K''}) \\
&= s_{K'}+s_{K''}+\frac{|D_{K'}||D_{K''}|}{|D_{K'}|+|D_{K''}|}(\bar{d}_{K'}-\bar{d}_{K''})^2.
\end{aligned}$$

Comparison methods

In order to evaluate the performance of our model, we compared our results to those of four existing methods. All of them require computing a dissimilarity measure between the connectivity patterns of elements i and j. For a connectivity matrix D,

$$W_{i,j} = \sqrt{\sum_{a\neq i,j}(D_{i,a}-D_{j,a})^2 + \sum_{a\neq i,j}(D_{a,i}-D_{a,j})^2}. \tag{3}$$

"Local similarity" computes the edge dissimilarity $W_{i,j}$ between each pair of neighboring elements, and then removes all edges above a given threshold. Here we set the threshold in order to obtain a desired number of clusters. This type of edge-finding approach has been used extensively for neuroimaging parcellation (*Cohen et al., 2008*; *Wig, Laumann & Petersen, 2014*; *Gordon et al., 2014*). Additionally, this is equivalent to using a spectral clustering approach (*Thirion et al., 2006*) if clustering in the embedding space is performed using single-linkage hierarchical clustering.

"Normalized cut" computes the edge similarity $S_{i,j} = 1/W_{i,j}$ between each pair of neighboring elements, then runs the normalized cut algorithm of *Shi & Malik (2000)*. This draws partitions between elements a and b when their edge similarity $S_{a,b}$ is low relative to their similarities with other neighbors. Although computing the globally optimal normalized cut is NP-complete, an approximate solution can be found quickly by solving a generalized eigenvalue problem. This approach has been specifically applied to neuroimaging data (*Craddock et al., 2012*).

"Region growing" is based on the approach described in *Blumensath et al. (2013)*. First, a set of seed points is selected which have high similarity to all their neighbors, since they are likely to be near the center of parcels. Seeds are then grown by iteratively adding neighboring elements with high similarity to the seed. Once every element has been assigned to a region, Ward clustering (see below) was used to cluster adjacent regions until the desired number of regions is reached.

"Ward clustering" requires computing $W_{i,j}$ between all pairs of elements (not just neighboring elements). Elements are each initialized as a separate cluster, and neighboring clusters are merged based on Ward's variance-minimizing linkage rule (*Ward, 1963*). This approach has been previously applied to neuroimaging data (*Thirion et al., 2014*; *Eickhoff et al., 2011*).

We also compared to random clusterings. Starting with each element in its own cluster, we iteratively picked a cluster uniformly at random and then merged it with a neighboring cluster (also picked uniformly at random from all neighbors). The process continued until the desired number of clusters remained.

Synthetic data

To generate synthetic connectivity data, we created three different parcellation patterns on an 18 by 18 grid (see Fig. 2), with the number of regions $K = 5,6,9$. Each element of the $K \times K$ connectivity matrix A was sampled from a standard normal distribution. For a given noise level σ, the connectivity value $D_{i,j}$ between element i in cluster $\mathbf{z_i}$ and element j in cluster $\mathbf{z_j}$ was sampled from a normal distribution with mean $A_{\mathbf{z_i},\mathbf{z_j}}$ and standard deviation σ. This data matrix was then input to our method with $\sigma_0^2 = 0.01$, which returned the MAP solution after 30 passes through the elements (approximately 10,000 steps). Both our method and all comparison methods were run for 20 different synthetic datasets for each noise level σ and the results were averaged.

We also performed a supplementary experiment using a more challenging three-spiral dataset (*Chang & Yeung, 2008*). We generated the connectivity matrix as above, and defined elements to be spatially adjacent if they were consecutive along a spiral or adjacent between neighboring spirals. In addition to our standard initialization scheme using the Ward clustering with highest probability according to our model, we also considered initializations with fixed numbers of clusters derived from Ward clustering ($K = 2,10$) or initializations in which the links c were chosen are random. The σ_0^2 hyperparameter was set to 0.01 as above, and the MAP solution was returned after 100 passes (or 1,000 passes for the random initialization).

Parcellations were evaluated by calculating their normalized mutual information (NMI) with the ground truth labeling. We calculate NMI as in *Strehl & Ghosh (2002)*. This measure ranges from 0 to 1, and does not require any explicit "matching" between parcels. For N total elements, if $\mathbf{z}$ assigns n_h elements to cluster h, $\mathbf{z_{gt}}$ assigns n_l^{gt} elements to cluster l, and $n_{h,l}$ elements are assigned to cluster h by $\mathbf{z}$ and cluster l by $\mathbf{z_{gt}}$, this is given by

$$\text{NMI}(\mathbf{z},\mathbf{z_{gt}}) = \frac{I(\mathbf{z},\mathbf{z_{gt}})}{\sqrt{H(\mathbf{z})H(\mathbf{z_{gt}})}} = \frac{\sum_h \sum_l n_{h,l} \log(N n_{h,l}/(n_h n_l^{gt}))}{\sqrt{\left(\sum_h n_h \log(n_h/N)\right)\left(\sum_l n_l^{gt} \log(n_l^{gt}/N)\right)}}. \quad (4)$$

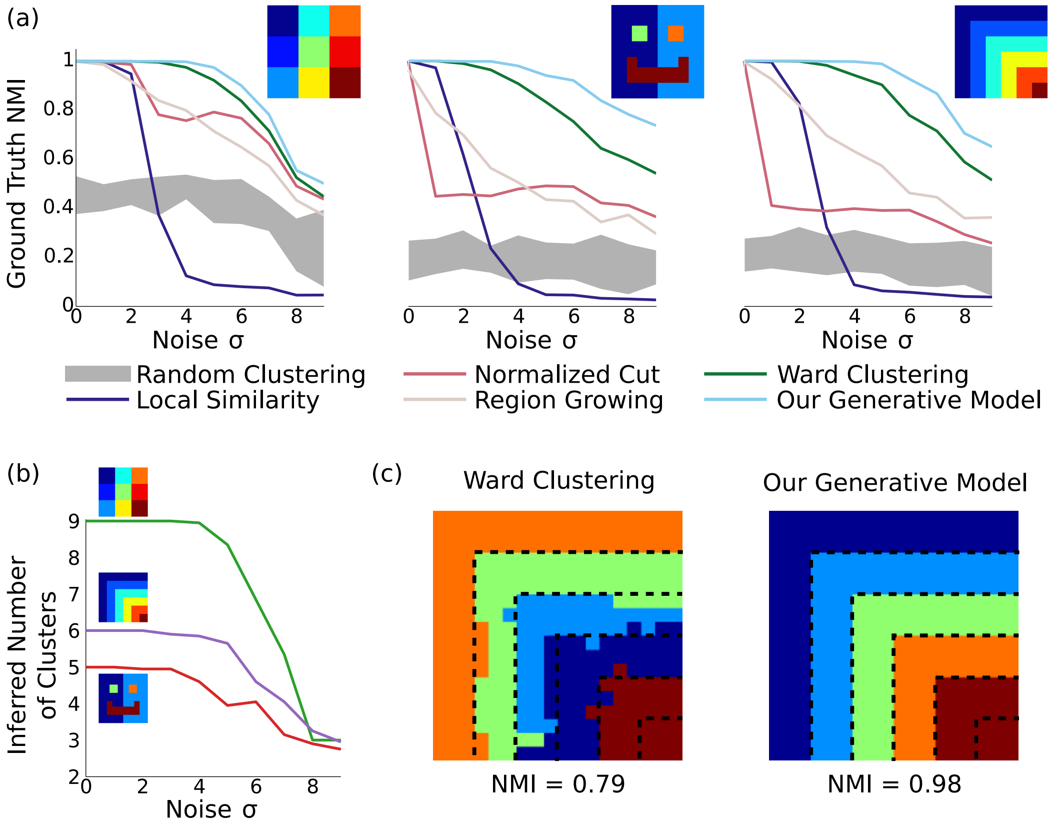

Figure 2 Results on synthetic data. (A) In three different synthetic datasets, our method is consistently better at recovering the ground-truth parcellation than alternative methods. This advantage is most pronounced when the parcels are arranged nonuniformly with unequal sizes, and the noise level is relatively high. Results are averaged across 20 random datasets for each noise level, and the gray region shows the standard deviation around random clusterings. (B) Our model can correctly infer the number of underlying clusters in the dataset for moderate levels of noise, and becomes more conservative about splitting elements into clusters as the noise level grows. (C) Example clusterings under the next-best clustering method and our model on the stripes dataset, for $\sigma = 6$. Although greedy clustering achieves a reasonable result, it is far noisier than the output of our method, which perfectly recovers the ground truth except for incorrectly merging the two smallest clusters.

Human brain functional data

We utilized group-averaged resting-state functional MRI correlation data from 468 subjects, provided by the Human Connectome Project's 500 Subjects release (*Van Essen et al., 2013*). Using a specialized Siemens 3T "Connectome Skyra" scanner (Siemens AG, Berlin, Germany), data was collected during four 15-min runs, during which subjects fixated with their eyes open on a small cross-hair. A multiband sequence was used, allowing for acquisition of 2.0 mm isotropic voxels at a rate of 720 ms. Data for each subject was cleaned using motion regression and ICA + FIX denoising (*Smith et al., 2013*; *Salimi-Khorshidi et al., 2014*) and then combined across subjects using an approximate group-PCA method yielding the strongest 4,500 spatial eigenvectors (*Smith et al., 2014*). The symmetric 59,412 by 59,412 functional connectivity matrix $D_{a,b}$ was computed as the correlation between the 4,500-dimensional eigenmaps of voxels a and b. For each of $\sigma_0^2 = 2{,}000, 3{,}000, 4{,}000, 5{,}000$, we ran Gibbs Sampling for 10 passes (approximately 600,000 steps) to find the MAP solution. For comparison with individual subjects, we also computed functional connectivity matrices for the first 20 subjects with resting-state data in the 500 Subjects release.

The map of retinotopic regions in visual cortex was created by mapping the volume-based atlas from (Wang et al., 2014) onto the Human Connectome group-averaged surface.

Human brain structural data

We obtained diffusion MRI data for 10 subjects from the Human Connectome Project's Q3 release (Van Essen et al., 2013). This data was collected on the specialized Skyra described above, using a multi-shell acquisition over 6 runs. Probabilistic tractgraphy was performed using FSL (Jenkinson et al., 2012), by estimating up to 3 crossing fibers with bedpostx (using gradient nonlinearities and a rician noise model) and then running probtrackx2 using the default parameters and distance correction. 2000 fibers were generated for each of the $1.7 \cdot 10^6$ white-matter voxels, yielding $3.4 \cdot 10^9$ total sampled tracks per subject (approximately 34 billion tracks in total). We assigned each of the endpoints to gray-matter voxels using the 32 k/hemisphere Conte69 registered standard mesh distributed for each subject, discarding the small number of tracks that did not have both endpoints in gray matter (e.g., cerebellar or spinal cord tracks). Since we are using distance correction, the weight of a track is set equal to its length. In order to account for imprecise tracking near the gray matter border, the weight of a track whose two endpoints are closest to voxels a and b is spread evenly across the connection between a and b, the connections between a and b's neighbors, and the connections between a's neighbors and b. Since the gray-matter mesh has a correspondence between subjects, we can compute the group-average number of tracks between every pair of voxels. Finally, since connectivity strengths are known to have a lognormal distribution (Markov et al., 2014), we define the symmetric 59,412 by 59,412 structural connectivity matrix $D_{a,b}$ as the log group-averaged weight between voxels a and b. The hyperparameter σ_0^2 was set to 3,000, and Gibbs Sampling was run for 10 passes (approximately 600,000 steps) to find the MAP solution.

Human migration data

We used the February 2014 release of the 2007–2011 county-to-county U.S. migration flows from the U.S. Census Bureau American Community Survey (ACS). This dataset includes estimates of the number of annual movers from every county to every other county, as well as population estimates for each county. We restricted our analysis to the continental U.S. To reduce the influence of noisy measurements from small counties, we preprocessed the dataset by iteratively merging the lowest-population county with its lowest-population neighbor (within the same state) until all regions contained at least 10,000 residents. This process produced 2,594 regions which we continue to refer to as "counties" for simplicity, though 306 cover multiple low-population counties. For visualization of counties and states, we utilized the KML Cartographic Boundary Files provided by the U.S. Census Bureau (KML).

One major issue with analyzing this migration data is that counties have widely varying populations (even after the preprocessing above), making it difficult to compare the absolute number of movers between counties. We correct for this by normalizing the migration flows relative to chance flows driven purely by population. If we assume a chance distribution in which a random mover is found to be moving from county a to county b based purely on population, then the normalized flow matrix is

$$D_{a,b} = \frac{M_{a,b}}{\left(\sum_{i,j} M_{i,j}\right) \cdot \frac{P_a P_b}{\left(\sum_i P_i\right)^2}} \tag{5}$$

where $M_{i,j}$ is the absolute number of movers from county i to county j, and P_i is the population of county i. This migration connectivity matrix D is therefore a nonnegative, asymmetric matrix in which values less than 1 indicate below-chance migration, and values greater than 1 indicate above-chance migration. Setting $\sigma_0^2 = 10$, we ran Gibbs Sampling for 50 passes (approximately 130,000 steps) to find the MAP solution.

RESULTS

Comparison on synthetic data

In order to understand the properties of our model and quantitatively compare it to alternatives on a dataset with a known ground truth, we performed several experiments with synthetic datasets. We compared against random parcellations (in which elements were randomly merged together) as well as four existing methods: local similarity, which simply thresholds the similarities between pairwise elements (similar to (*Thirion et al., 2006*; *Cohen et al., 2008*; *Wig, Laumann & Petersen, 2014*; *Gordon et al., 2014*)); normalized cut (*Craddock et al., 2012*) which finds parcels maximizing within-cluster similarity and between-cluster difference; region growing (*Blumensath et al., 2013*), an agglomerative clustering method which selects stable points and iteratively merges similar elements; and Ward clustering (*Thirion et al., 2014*), an agglomerative clustering method which iteratively merges elements to minimize the total variance. Since these methods cannot automatically discover the number of clusters, they (and the random clustering) are set to use the same number of clusters as inferred by our method. We varied the noise level of the synthetic connectivity matrix from low to high, and evaluated the learned clusters using the normalized mutual information with the ground truth, which ranges from 0 to 1 (with 1 indicating perfect recovery).

As shown in Fig. 2, our method identifies parcels that best match the ground truth, across all three datasets and all noise levels. The naive local similarity approach performs very poorly under even mild noise conditions, and becomes worse than chance for high noise levels (for which most parcellations consist of single noisy voxels). Normalized cut is competitive only when the ground-truth parcels are equally sized (matching results from (*Blumensath et al., 2013*)), and is near-chance in the other cases. Region growing is more consistent across datasets, but does not reach the performance of Ward clustering, which outperforms all methods other than ours. Our model correctly infers the number of clusters with moderate amounts of noise (using the same hyperparameters in all experiments), and finds near-perfect parcellations even at very high noise levels (see Fig. 2C).

We also evaluated our model on a three-spiral dataset previously used in clustering work (*Chang & Yeung, 2008*), showing that we outperform other methods regardless of initialization scheme (see Figure S1).

Functional connectivity in the human brain

To investigate the spatial structure of functional connectivity in the human brain, we applied our model to data from the Human Connectome Project (*Van Essen et al., 2013*). Combining data from 468 subjects, this symmetric 59,412 by 59,412 matrix gives the correlation between fMRI timecourses of every pair of vertices on the surface of the brain (at 2 mm resolution) during a resting-state scan (in which subjects fixated on a blank screen). Using the anatomical surface models provided with the data, we defined vertices to be spatially adjacent if they were neighbors along the cortical surface.

Evaluating cortical parcellations is challenging since there is no clear ground truth for comparison, and different applications could require parcellations with different types of properties (e.g., optimizing for fitting individual subjects or for stability across subjects (*Thirion et al., 2014*)). One simple measure of an effective clustering is the fraction of variance in the full 3.5 billion element matrix which is captured by the connectivity between parcels (consisting of only tens of thousands of connections). As shown in Fig. 3A, our parcellation explains more variance for a given number of clusters than greedy Ward clustering; in order to achieve the same level of performance as our model, the simpler approach would require approximately 30 additional clusters. We can also measure how well this group-level parcellation (using data averaged from hundreds of subjects) fits the data from 20 individual subjects. Although the variance explained is substantially smaller for individual subjects, due both to higher noise levels and inter-subject connectivity differences, our model explains significantly more variance than Ward clustering with 140 clusters ($t_{19} = 2.97, p < 0.01$ one-tailed t-test), 155 clusters ($t_{19} = 3.67, p < 0.01$), or 172 clusters ($t_{19} = 1.77, p < 0.05$). The 220-cluster solutions from our model and Ward clustering generalize equally well, suggesting that our method's largest gains over greedy approximation occur in the more challenging regime of small numbers of clusters.

One part of the brain in which we do have prior knowledge about cortical organization is in visual cortex, which is segmented into well-known retinotopic field maps (*Wang et al., 2014*). We can qualitatively examine the match between our 172-cluster parcellation (Fig. 3C) and these retinotopic maps on an inflated cortical surface, shown in Fig. 3D. First, we observe a wide variety in the size and shape of the learned parcels, since the model places no explicit constraints on the clusters except that they must be spatially connected. We also see that we correctly infer very similar parcellations between hemispheres, despite the fact that bilateral symmetry is not enforced by the model. The earliest visual field maps (V1, V2, V3, hV4, LO1, LO2) all radiate out from a common representation of the fovea (*Brewer & Barton, 2012*), and in this region, our model generates ring parcellations

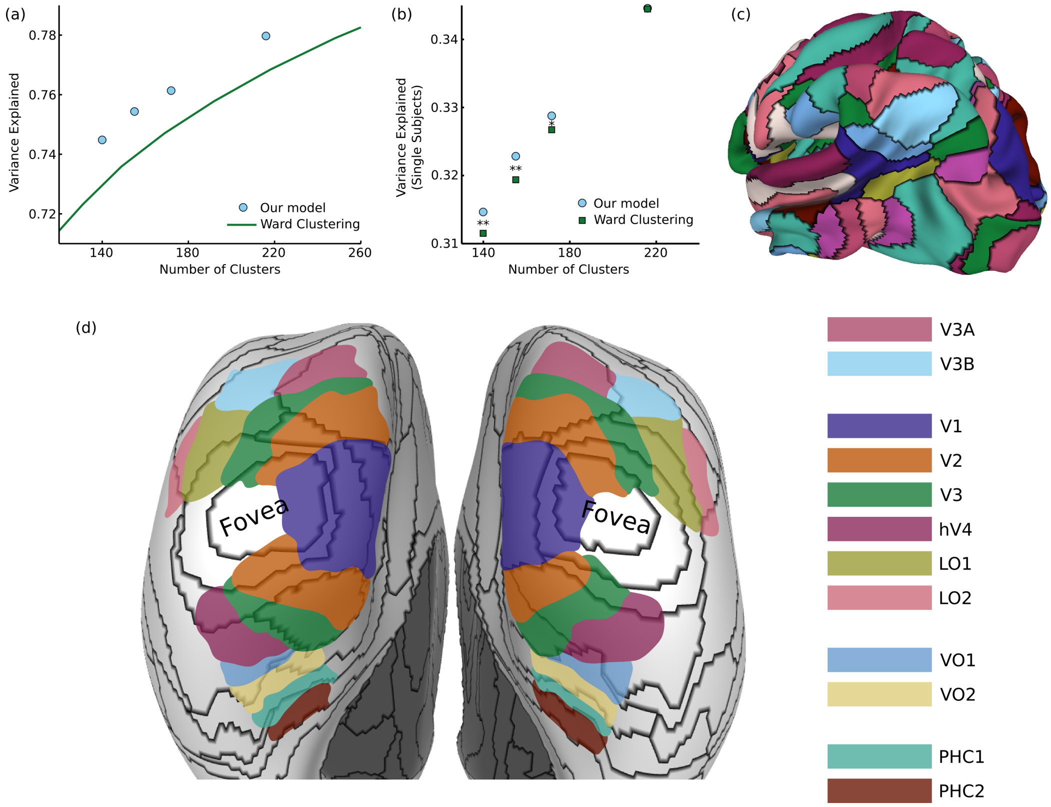

Figure 3 Results on functional brain connectivity (A) Our model consistently provides a better fit to the data than greedy clustering, explaining the same amount of variance with 30 fewer clusters (different points were generated from different values of the hyperparameter σ_0^2). (B) When using our group-learned clustering to explain variance in 20 individual subjects, we consistently generalize better than the greedy clusters for cluster sizes less than 200 (* $p < 0.05$, ** $p < 0.01$). (C) A sample 172-cluster parcellation from our method. (D) Comparison between our parcels and retinotopic maps, showing a transition from eccentricity-based divisions to field map divisions.

which divide the visual field based on distance from the fovea. The parcellation also draws a sharp border between peripheral V1 and V2. In the dorsal V3A/V3B cluster, V3A and V3B are divided into separate parcels. In medial temporal regions, parcel borders show an approximate correspondence with known VO and PHC borders, with an especially close match along the PHC1-PHC2 border. Overall, we therefore see a transition from an eccentricity-based parcellation in the early visual cluster to a parcellation corresponding to known field maps in the later dorsal and ventral visual areas.

Structural connectivity in the human brain

Based on diffusion MRI data from the Human Connectome Project (*Van Essen et al., 2013*), we used probabilistic tractography (*Behrens et al., 2007*) to generate estimates of the strength of the structural fiber connections between each pair of 2 mm gray-matter voxels. Approximately 34 billion tracts were sampled across 10 subjects, yielding a symmetric 59,412 by 59,412 matrix in which about two-thirds of the elements are non-zero. Applying our method to this matrix parcellates the brain into groups of voxels that all had the

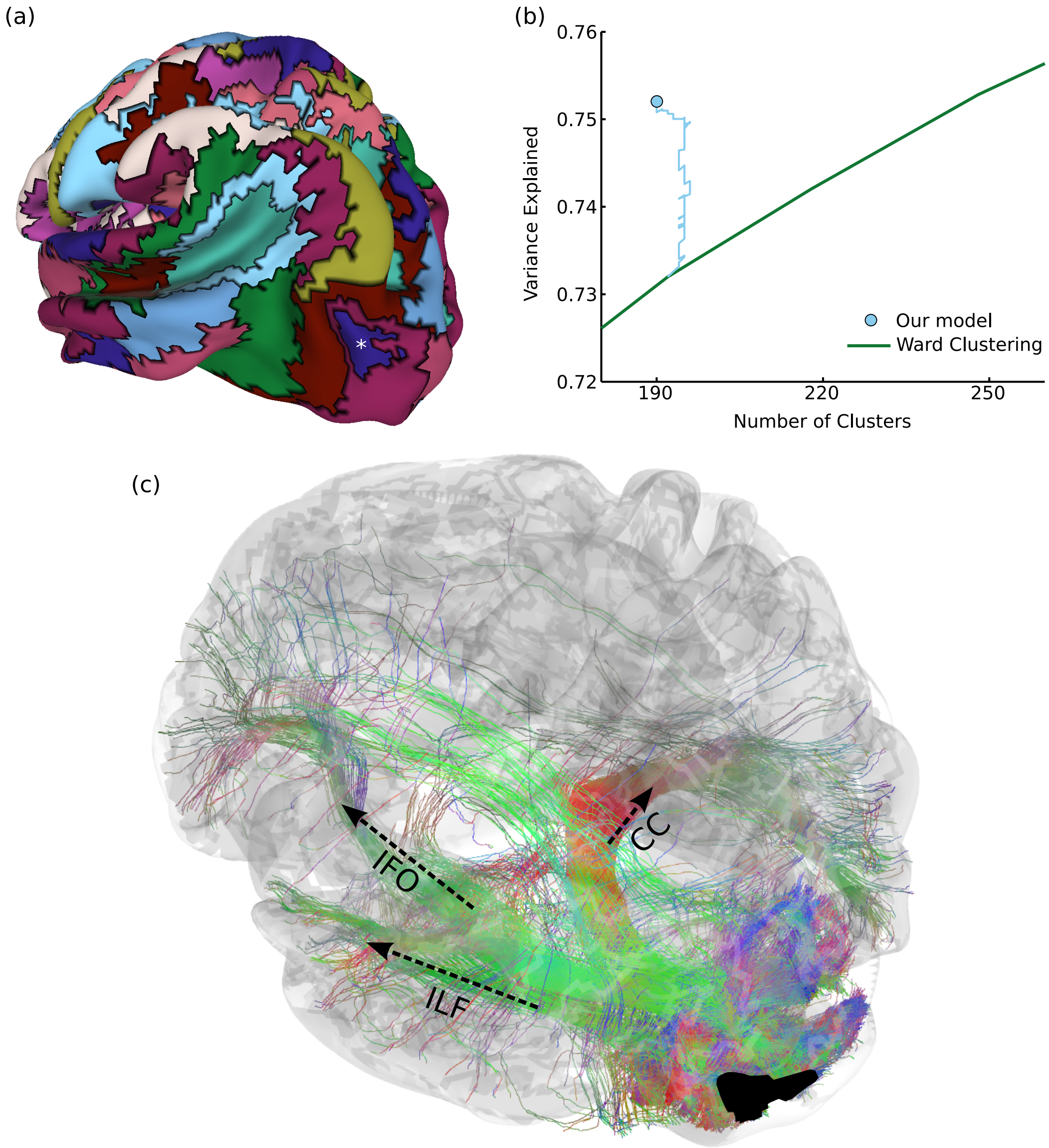

Figure 4 Results on structural brain connectivity. (A) A 190-cluster parcellation of the brain based on structural tractography patterns. (B) This parcellation fits the data substantially better than greedy clustering, which would require an additional 55 clusters to explain the same amount of variance. The blue path shows how our model fit improves over the course of Gibbs sampling when initialized with the greedy solution. (C) An example of 35,000 tracks (from one subject) connected to a parcel in the lateral occipital sulcus, marked with an asterisk in (A). These include portions of major fascicles such as the inferior longitudinal fasciculus (ILF), inferior fronto-occipital fasciculus (IFO), and corpus callosum (CC).

same distribution of incident fibers. This problem is even more challenging than in the functional case, since this matrix is much less spatially smooth.

Figure 4A shows a 190-region parcellation. Our clustering outperforms greedy clustering by an even larger margin than with the functional data, explaining as much variance as a greedy parcellation with 55 additional clusters. Figure 4B also shows how the model fit evolves over many rounds of Gibbs sampling, when initialized with the greedy solution. Since our method can flexibly explore different numbers of clusters, it is able (unlike a greedy method) to perform complex splitting and merging operations on the parcels. Qualitatively evaluating our parcellation is even more challenging than in the previous functional experiment, but we find that our parcels match the endpoints of major known tracts. For example, Fig. 4C shows 35,000 probabilistically-sampled tracts intersecting with a parcel in the left lateral occipital sulcus, which (in addition to many short-range fibers) connects to the temporal lobe through the inferior longitudinal fasciculus, to the frontal lobe through the inferior fronto-occipital fasciculus, and to homologous regions in the right hemisphere through the corpus callosum (*Wakana et al., 2004*). Note that the full connectivity matrix was constructed from a million times as many tracks as shown in this figure, in order to estimate the pairwise connectivity between every pair of gray-matter voxels.

Human migration in the United States

Given our successful results on neuroimaging data, we then applied our method to an entirely distinct dataset: internal migration within the United States. Using our probabilistic model, we sought to summarize the (asymmetric) matrix of migration between US counties as flows between a smaller number of contiguous regions. The model is essentially searching for a parcellation such that all counties within a parcel have similar (in- and out-) migration patterns. Note that this is a challenging dataset for clustering analyses since the county-level migration matrix is extremely noisy and sparse, with only 3.8% of flows having a nonzero value.

As shown in Fig. 5A, we identify 83 regions defined by their migration properties. There are a number of interesting properties of this parcellation of the United States. Many clusters share borders with state borders, even though no information about the state membership of different counties was used during the parcellation. This alignment was substantially more prominent than when generating random 83-cluster parcellations, as shown in Fig. 5B. As described in the Discussion, this is consistent with previous work showing behavioral differences caused by state borders, providing the first evidence that state membership also has an impact on intranational migration patterns. Greedy clustering performs very poorly on this sparse, noisy matrix, producing many clusters containing only one or a small number of counties, and has a lower NMI with state borders than even the random parcellations.

The 10 most populous clusters (Fig. 5C) cover 18 of the 20 largest cities in the US, with the two largest parcels covering the Northeast and the west coast. Some clusters roughly align with states or groups of states, while other divide states (e.g., the urban centers of east Texas) or cut across multiple states (e.g., the “urban midwest” cluster consisting of Columbus, Detroit, and Chicago). As shown in Fig. 5D, our method succeeds in reordering the migration matrix to be composed of approximately piecewise constant blocks. In this case (and in many applications) the blocks along the main diagonal are most prominent, but this assortative structure is not enforced by the model. Though largely symmetric,

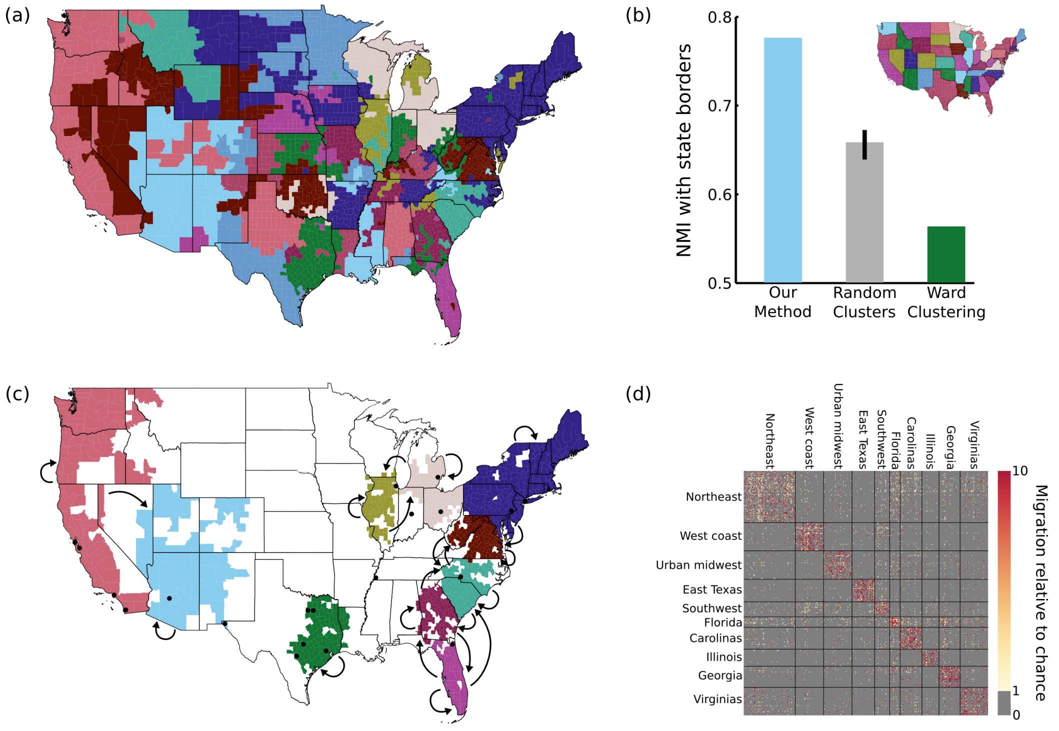

Figure 5 Results on migration dataset. (A) Our parcellation identified 83 contiguous regions within the continental US, such that migration between these regions summarizes the migration between all 2594 counties. (B) This parcellation was better aligned with state borders than an 83-cluster random parcellation (95% confidence interval shown) or an 83-cluster greedy Ward parcellation. (C) The top 10 clusters (by population) are shown, with arrows indicating above-chance flows between the clusters. The 20 most populous US cities are indicated with black dots for reference. (D) A portion of the migration matrix, showing the 1051 counties covered by the top 10 clusters.

some flows do show large asymmetries. For example, the two most asymmetrical flows by absolute difference are between the urban midwest and Illinois (out of Illinois = 1.3, into Illinois = 2.0), and Florida and Georgia (out of Georgia = 1.3, into Georgia = 2.0).

DISCUSSION

In this work we have introduced a new generative nonparametric model for parcellating a spatial map based on connectivity information. After showing that our model outperforms existing baselines on synthetic data, we applied it to three distinct real-world datasets: functional brain connectivity, structural brain connectivity, and US migration. In each case our method showed improvements over the current state-of-the-art, and was able to capture hidden spatial patterns in the connectivity data. The gap between our approach and past work varied with the difficulty of the parcellation problem; hierarchical clustering would require $\sim 17\%$ more clusters for the relatively smooth functional connectivity data and $\sim 29\%$ more clusters for the more challenging structural connectivity data, and fails completely for the most noisy migration dataset.

Finding a connectivity-based parcellation of the brain's cortical surface has been an important goal in recent neuroimaging research, for two primary reasons. First, the shapes and locations of connectivity-defined regions may help inform us about underlying modularity in cortex, providing a relatively hypothesis-free delineation of regions with distinct functional or structural properties. For example, connectivity clustering has been used to identify substructures in the posterior medial cortex (*Bzdok et al., 2014*), temporoparietal junction (*Mars et al., 2012*), medial frontal cortex (*Johansen-Berg et al., 2004*; *Kim et al., 2010*; *Crippa et al., 2011*; *Klein et al., 2007*), occipital lobes (*Thiebaut de Schotten et al., 2014*), frontal pole (*Moayedi et al., 2014*; *Liu et al., 2013*), lateral premotor cortex (*Tomassini et al., 2007*), lateral parietal cortex (*Mars et al., 2011*; *Ruschel et al., 2013*), amygdala (*Cheng & Fan, 2014*; *Mishra et al., 2014*), and insula (*Cauda et al., 2011*). Second, an accurate parcellation is necessary for performing higher-level analysis, such as analyzing distributed connectivity networks among parcels (*Power et al., 2013*; *Andrews-Hanna et al., 2010*; *Van den Heuvel & Sporns, 2013*), using connectivity as a clinical biomarker (*Castellanos et al., 2013*), or pooling voxel features for classification (*Xu, Zhen & Liu, 2010*). Consistent with our results, previous work has found that greedy Ward clustering generally fits the datasets best (in terms of variance explained) among these existing methods (*Thirion et al., 2014*).

Our finding of eccentricity-based resting-state parcels in early visual areas is consistent with previous results showing a foveal vs. peripheral division of visual regions based on connectivity (*Thomas Yeo et al., 2011*; *Lee et al., 2012*). Since our parcellation is much higher-resolution, we are able to observe nested clusters at multiple eccentricities. Our results are the first to suggest that higher-level retinotopic regions, especially PHC1 and PHC2, have borders that are related to changes in connectivity properties.

Parcellation based on structural tractography has generally been limited to specific regions of interest (*Mars et al., 2012*; *Johansen-Berg et al., 2004*; *Crippa et al., 2011*; *Klein et al., 2007*; *Thiebaut de Schotten et al., 2014*; *Moayedi et al., 2014*; *Liu et al., 2013*; *Tomassini et al., 2007*; *Mars et al., 2011*; *Ruschel et al., 2013*), in part due to the computational difficulties of computing and analyzing a full voxel-by-voxel connectivity matrix. Our parcellation for this modality is somewhat preliminary; probabilistic tractography algorithms are still in their infancy, with recent work showing that they produce many tracts that are not well-supported by the underlying diffusion data (*Pestilli et al., 2014*) and are of questionable anatomical accuracy (*Thomas et al., 2014*). As diffusion imaging and tractography methods continue to improve, the input connectivity matrix to our method will become higher quality and allow for more precise parcellation.

There has been detailed scientific study of both inter- and intra-national migration patterns for over a century, beginning with the 1885 work of *Ravenstein (1885)*. Even in this initial study (within the UK), it was clear that migration properties varied with spatial location; for example, rural areas showed large out-migration, while metropolitan areas showed greater in-migration, including long-distance migrants. The impact of state borders on migration behavior has not, to our knowledge, been specifically addressed, but there is a growing literature documenting differences in behaviors across state lines. Neighboring counties across state lines are less politically similar than those within a state, suggesting that a state border "creates a barrier to, or contains, political and economic institutions, policies, and possibly movement" (*Tam Cho & Nicley, 2008*). State borders also play a role in isolating communities economically; this phenomenon gained a great of attention after Wolf's 2000 study (*Wolf, 2000*), showing that trade was markedly lower between states than within states (controlling for distance using a gravity model). Our results demonstrate in a hypothesis-free way that migration behavior is influenced by state identities, since our method discovers a parcellation related in many regions to state borders, without being given any information about the state membership of each county. Our results also show that state borders alone are not sufficient to capture the complexities of migration behavior, since other factors can override state identities to create other types of communities (such as in our "Urban midwest" parcel).

Since our algorithm makes many passes over the dataset, it does take longer than previous methods to find the most likely clustering. There are a number of possible approaches for speeding up inference which could be explored in future work. One possibility is to parallelize inference by performing Gibbs sampling on multiple elements simultaneously; although this would no longer be guaranteed to converge to the true posterior distribution, in practice this may not be an issue. Another option is to compute the Gibbs sampling probabilities only approximately (*Korattikara, Chen & Welling, 2014*), by using only a random subset of connectivities in a large matrix to approximate the likelihood of a proposed parcellation. It also may be possible to increase the performance of our algorithm even further by starting with many different initializations and selecting the solution with highest MAP probability.

CONCLUSIONS

In summary, we have proposed the first general-purpose probabilistic model to intrinsically incorporate spatial information in its clustering prior, allowing us to search directly in the space of contiguous parcellations using collapsed Gibbs sampling. Our approach is far more flexible and precise than previous work, with no constraints on the sizes and shapes of the learned parcels. This makes our model more resilient to noise in synthetic tests, and provides better fits to real-world data drawn from three different domains. This diverse set of results suggests that our model could be applied to a large set of biological network datasets to reveal fine-grained structure in spatial maps. We have publicly released both MATLAB and python implementations of our method at http://goo.gl/xys4xh under a BSD open-source licence.

ACKNOWLEDGEMENTS

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Thank you Henry Jung for porting our MATLAB code to python, to Mike Arcaro for providing the map of retinotopic regions in visual cortex, and to Michelle Greene for reviewing early versions of this draft.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Funding was provided by a National Science Foundation Graduate Research Fellowship under grant number DGE-0645962, and the Office of Naval Research Multidisciplinary University Research Initiative grant number N000141410671. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Science Foundation Graduate Research Fellowship: DGE-0645962.
Office of Naval Research Multidisciplinary University Research Initiative: N000141410671.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Christopher Baldassano conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Diane M. Beck and Li Fei-Fei conceived and designed the experiments, wrote the paper, reviewed drafts of the paper.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.784#supplemental-information.

REFERENCES

Abraham A, Dohmatob E, Thirion B, Samaras D, Varoquaux G. 2013. Extracting brain regions from rest fMRI with total-variation constrained dictionary learning. In: Mori K, Sakuma I, Sato Y, Barillot C, Navab N, eds. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*. Berlin: Springer-Verlag, 607–615 DOI 10.1007/978-3-642-40763-5_75.

Aicher C, Jacobs AZ, Clauset A. 2014. Learning latent block structure in weighted networks. *Journal of Complex Networks* (online) DOI 10.1093/comnet/cnu026.

Alon U. 2003. Biological networks: the tinkerer as an engineer. *Science* **301(5641)**:1866–1867 DOI 10.1126/science.1089072.

Andrews-Hanna JR, Reidler JS, Sepulcre J, Poulin R, Buckner RL. 2010. Functional-anatomic fractionation of the brain's default network. *Neuron* **65(4)**:550–562 DOI 10.1016/j.neuron.2010.02.005.

Barabási A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nature Reviews. Genetics* **5(2)**:101–113 DOI 10.1038/nrg1272.

Behrens TE, Berg HJ, Jbabdi S, Rushworth MF, Woolrich MW. 2007. Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage* **34(1)**:144–155 DOI 10.1016/j.neuroimage.2006.09.018.

Blei DM, Frazier PI. 2011. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research* **12**:2461–2488.

Blumensath T, Jbabdi S, Glasser MF, Van Essen DC, Ugurbil K, Behrens TE, Smith SM. 2013. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. *Neuroimage* **76**:313–324 DOI 10.1016/j.neuroimage.2013.03.024.

Brewer AA, Barton B. 2012. Visual field map organization in human visual cortex. In: Molotchnikoff S, Rouat J, eds. *Visual cortex-current status and perspectives*. Rijeka: InTech, 29–60 DOI 10.5772/51914.

Bzdok D, Heeger A, Langner R, Laird AR, Fox PT, Palomero-Gallagher N, Vogt BA, Zilles K, Eickhoff SB. 2014. Subspecialization in the human posterior medial cortex. *Neuroimage* **106C**:55–71.

Castellanos FX, Di Martino A, Craddock RC, Mehta AD, Milham MP. 2013. Clinical applications of the functional connectome. *NeuroImage* **80**:527–540 DOI 10.1016/j.neuroimage.2013.04.083.

Cauda F, D'Agata F, Sacco K, Duca S, Geminiani G, Vercelli A. 2011. Functional connectivity of the insula in the resting brain. *Neuroimage* **55(1)**:8–23 DOI 10.1016/j.neuroimage.2010.11.049.

Chang H, Yeung D-Y. 2008. Robust path-based spectral clustering. *Pattern Recognition* **41(1)**:191–203 DOI 10.1016/j.patcog.2007.04.010.

Cheng H, Fan Y. 2014. Semi-supervised clustering for parcellating brain regions based on resting state fMRI data. In: Ourselin S, Styner MA, eds. *Proceedings of SPIE 9034, Medical Imaging 2014: Image Processing 903427*. DOI 10.1117/12.2043467.

Cohen AL, Fair DA, Dosenbach NUF, Miezin FM, Dierker D, Van Essen DC, Schlaggar BL, Petersen SE. 2008. Defining functional areas in individual human brains using resting functional connectivity MRI. *NeuroImage* **41(1)**:45–57 DOI 10.1016/j.neuroimage.2008.01.066.

Cordes D, Haughton V, Carew JD, Arfanakis K, Maravilla K. 2002. Hierarchical clustering to measure connectivity in fMRI resting-state data. *Magnetic Resonance Imaging* **20(4)**:305–317 DOI 10.1016/S0730-725X(02)00503-9.

Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS. 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping* **33(8)**:1914–1928 DOI 10.1002/hbm.21333.

Crippa A, Cerliani L, Nanetti L, Roerdink JBTM. 2011. Heuristics for connectivity-based brain parcellation of SMA/pre-SMA through force-directed graph layout. *NeuroImage* **54(3)**:2176–2184 DOI 10.1016/j.neuroimage.2010.09.075.

Eickhoff SB, Bzdok D, Laird AR, Roski C, Caspers S, Zilles K, Fox PT. 2011. Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. *Neuroimage* **57(3)**:938–949 DOI 10.1016/j.neuroimage.2011.05.021.

Freilich S, Kreimer A, Meilijson I, Gophna U, Sharan R, Ruppin E. 2010. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Research* **38(12)**:3857–3868 DOI 10.1093/nar/gkq118.

Golland Y, Golland P, Bentin S, Malach R. 2008. Data-driven clustering reveals a fundamental subdivision of the human cortex into two global systems. *Neuropsychologia* **46(2)**:540–553 DOI 10.1016/j.neuropsychologia.2007.10.003.

Golland P, Golland Y, Malach R. 2007. Detection of spatial activation patterns as unsupervised segmentation of fMRI data. *Medical Image Computing and Computer-Assisted Intervention* **10(Pt 1)**:110–118.

Gorbach NS, Schütte C, Melzer C, Goldau M, Sujazow O, Jitsev J, Douglas T, Tittgemeyer M. 2011. Hierarchical information-based clustering for connectivity-based cortex parcellation. *Frontiers in Neuroinformatics* **5**:18 DOI 10.3389/fninf.2011.00018.

Gordon EM, Laumann TO, Adeyemo B, Huckins JF, Kelley WM, Petersen SE. 2014. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral Cortex* Epub ahead of print DOI 10.1093/cercor/bhu239.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. *Nature* **402(December)**:47–52 DOI 10.1038/35011540.

Heller R, Stanley D, Yekutieli D, Rubin N, Benjamini Y. 2006. Cluster-based analysis of FMRI data. *Neuroimage* **33(2)**:599–608 DOI 10.1016/j.neuroimage.2006.04.233.

Honnorat N, Eavani H, Satterthwaite TD, Gur RE, Gur RC, Davatzikos C. 2014. GraSP: geodesic graph-based segmentation with shape priors for the functional parcellation of the cortex. *Neuroimage* **106**:207–221 DOI 10.1016/j.neuroimage.2014.11.008.

Jbabdi S, Woolrich MW, Behrens TE. 2009. Multiple-subjects connectivity-based parcellation using hierarchical Dirichlet process mixture models. *Neuroimage* **44(2)**:373–384 DOI 10.1016/j.neuroimage.2008.08.044.

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. 2012. FSL. *Neuroimage* **62(2)**:782–790 DOI 10.1016/j.neuroimage.2011.09.015.

Johansen-Berg H, Behrens TEJ, Robson MD, Drobnjak I, Rushworth MFS, Brady JM, Smith SM, Higham DJ, Matthews PM. 2004. Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America* **101(36)**:13335–13340 DOI 10.1073/pnas.0403743101.

Kim JH, Lee JM, Jo HJ, Kim SH, Lee JH, Kim ST, Seo SW, Cox RW, Na DL, Kim SI, Saad ZS. 2010. Defining functional SMA and pre-SMA subregions in human MFC using resting state fMRI: functional connectivity-based parcellation method. *Neuroimage* **49(3)**:2375–2386 DOI 10.1016/j.neuroimage.2009.10.016.

Klein JC, Behrens TEJ, Robson MD, Mackay CE, Higham DJ, Johansen-Berg H. 2007. Connectivity-based parcellation of human cortex using diffusion MRI: establishing reproducibility, validity and observer independence in BA 44/45 and SMA/pre-SMA. *NeuroImage* **34(1)**:204–211 DOI 10.1016/j.neuroimage.2006.08.022.

KML–Cartographic Boundary Files—Geography—U.S. Census Bureau. *Available at http://www.census.gov/geo/maps-data/data/tiger-kml.html* (accessed 17 April 2014).

Korattikara A, Chen Y, Welling M. 2014. Austerity in MCMC Land: cutting the Metropolis–Hastings Budget. In: *Proceedings of the 31st international conference on machine learning. Available at http://machinelearning.wustl.edu/mlpapers/paper_files/icml2014c1_korattikara14.pdf*.

Krause AE, Frank KA, Mason DM. 2003. Compartments revealed in food-web structure. *Nature* **426**:282–285 DOI 10.1038/nature02115.

Lee MH, Hacker CD, Snyder AZ, Corbetta M, Zhang D, Leuthardt EC, Shimony JS. 2012. Clustering of resting state networks. *PLoS ONE* **7(7)**:e40370 DOI 10.1371/journal.pone.0040370.

Legendre P, Fortin MJ. 1989. Spatial pattern and ecological analysis. *Vegetatio* **80**:107–138 DOI 10.1007/BF00048036.

Liu H, Qin W, Li W, Fan L, Wang J, Jiang T, Yu C. 2013. Connectivity-based parcellation of the human frontal pole with diffusion tensor imaging. *The Journal of Neuroscience* **33(16)**:6782–6790 DOI 10.1523/JNEUROSCI.4882-12.2013.

Markov NT, Ercsey-Ravasz MM, Ribeiro Gomes AR, Lamy C, Magrou L, Vezoli J, Misery P, Falchier A, Quilodran R, Gariel MA, Sallet J, Gamanut R, Huissoud C, Clavagnier S, Giroud P, Sappey-Marinier D, Barone P, Dehay C, Toroczkai Z, Knoblauch K, Van Essen DC, Kennedy H. 2014. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral Cortex* **24(1)**:17–36 DOI 10.1093/cercor/bhs270.

Mars RB, Jbabdi S, Sallet J, O'Reilly JX, Croxson PL, Olivier E, Noonan MP, Bergmann C, Mitchell AS, Baxter MG, Behrens TEJ, Johansen-Berg H, Tomassini V, Miller KL, Rushworth MFS. 2011. Diffusion-weighted imaging tractography-based parcellation of the human parietal cortex and comparison with human and macaque resting-state functional connectivity. *The Journal of Neuroscience* **31(11)**:4087–4100 DOI 10.1523/JNEUROSCI.5102-10.2011.

Mars RB, Sallet J, Schüffelgen U, Jbabdi S, Toni I, Rushworth MFS. 2012. Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. *Cerebral Cortex* **22(8)**:1894–1903 DOI 10.1093/cercor/bhr268.

Migration/Geographic Mobility—County-to-County Migration Flows. 2007–2011. ACS—People and Households—U.S. Census Bureau. *Available at http://www.census.gov/hhes/migration/data/acs/county_to_county_mig_2007_to_2011.html* (accessed 6 February 2014).

Mishra A, Rogers BP, Chen LM, Gore JC. 2014. Functional connectivity-based parcellation of amygdala using self-organized mapping: a data driven approach. *Human Brain Mapping* **35(4)**:1247–1260 DOI 10.1002/hbm.22249.

Moayedi M, Salomons TV, Dunlop KA, Downar J, Davis KD. 2014. Connectivity-based parcellation of the human frontal polar cortex. *Brain Structure & Function* Epub ahead of print DOI 10.1007/s00429-014-0809-6.

Moreno-Dominguez D, Anwander A, Knosche TR. 2014. A hierarchical method for whole-brain connectivity-based parcellation. *Human Brain Mapping* **35(10)**:5000–5025 DOI 10.1002/hbm.22528.

Morup M, Madsen K, Dogonowski AM, Siebner H, Hansen LK. 2010. Infinite relational modeling of functional connectivity in resting state fmri. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, eds. *Advances in neural information processing systems 23*. 1750–1758. *Available at http://papers.nips.cc/paper/4057-infinite-relational-modeling-of-functional-connectivity-in-resting-state-fmri.pdf*.

Mumford JA, Horvath S, Oldham MC, Langfelder P, Geschwind DH, Poldrack RA. 2010. Detecting network modules in fMRI time series: a weighted network analysis approach. *Neuroimage* **52(4)**:1465–1476 DOI 10.1016/j.neuroimage.2010.05.047.

Murphy KP. 2007. Conjugate bayesian analysis of the gaussian distribution. Technical report. UBC.

Olesen JM, Bascompte J, Dupont YL, Jordano P. 2007. The modularity of pollination networks. *Proceedings of the National Academy of Sciences of the United States of America* **104(50)**:19891–19896 DOI 10.1073/pnas.0706375104.

Pestilli F, Yeatman JD, Rokem A, Kay KN, Wandell BA. 2014. Evaluation and statistical inference for human connectomes. *Nature Methods* **11(10)**:1058–1063 DOI 10.1038/nmeth.3098.

Power JD, Schlaggar BL, Lessov-Schlaggar CN, Petersen SE. 2013. Evidence for hubs in human functional brain networks. *Neuron* **79**(**4**):798–813 DOI 10.1016/j.neuron.2013.07.035.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. 2002. Hierarchical organization of modularity in metabolic networks.. *Science* **297**(**5586**):1551–1555 DOI 10.1126/science.1073374.

Ravenstein EG. 1885. The laws of migration. *Journal of the Statistical Society of London* **48**(**2**):167–235 DOI 10.2307/2979181.

Rives AW, Galitski T. 2003. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America* **100**(**3**):1128–1133 DOI 10.1073/pnas.0237338100.

Ruschel M, Knösche TR, Friederici AD, Turner R, Geyer S, Anwander A. 2013. Connectivity architecture and subdivision of the human inferior parietal cortex revealed by diffusion MRI. *Cerebral Cortex* **24**(**9**):2436–2448 DOI 10.1093/cercor/bht098.

Ryali S, Chen T, Supekar K, Menon V. 2013. A parcellation scheme based on von Mises-Fisher distributions and Markov random fields for segmenting brain regions using resting-state fMRI. *NeuroImage* **65**:83–96 DOI 10.1016/j.neuroimage.2012.09.067.

Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM. 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* **90**:449–468 DOI 10.1016/j.neuroimage.2013.11.046.

Shi J, Malik J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(**8**):888–905 DOI 10.1109/34.868688.

Smith SM, Beckmann CF, Andersson J, Auerbach EJ, Bijsterbosch J, Douaud G, Duff E, Feinberg DA, Griffanti L, Harms MP, Kelly M, Laumann T, Miller KL, Moeller S, Petersen S, Power J, Salimi-Khorshidi G, Snyder AZ, Vu AT, Woolrich MW, Xu J, Yacoub E, Ugurbil K, Van Essen DC, Glasser MF. 2013. Resting-state fMRI in the human connectome project. *Neuroimage* **80**:144–168 DOI 10.1016/j.neuroimage.2013.05.039.

Smith SM, Hyvarinen A, Varoquaux G, Miller KL, Beckmann CF. 2014. Group-PCA for very large fMRI datasets. *Neuroimage* **101**:738–749 DOI 10.1016/j.neuroimage.2014.07.051.

Smith SM, Miller KL, Salimi-Khorshidi G, Webster M, Beckmann CF, Nichols TE, Ramsey JD, Woolrich MW. 2011. Network modelling methods for FMRI. *NeuroImage* **54**(**2**):875–891 DOI 10.1016/j.neuroimage.2010.08.063.

Strehl A, Ghosh J. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* **3**:583–617.

Tam Cho WK, Nicley EP. 2008. Geographic proximity versus institutions: evaluating borders as real political boundaries. *American Politics Research* **36**(**6**):803–823 DOI 10.1177/1532673X08316701.

Thiebaut de Schotten M, Urbanski M, Valabregue R, Bayle DJ, Volle E. 2014. Subdivision of the occipital lobes: an anatomical and functional MRI connectivity study. *Cortex* **56**:121–137 DOI 10.1016/j.cortex.2012.12.007.

Thirion B, Flandin G, Pinel P, Roche A, Ciuciu P, Poline JB. 2006. Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. *Human Brain Mapping* **27**(**8**):678–693 DOI 10.1002/hbm.20210.

Thirion B, Varoquaux G, Dohmatob E, Poline JB. 2014. Which fMRI clustering gives good brain parcellations? *Frontiers in Neuroscience* **8**:167 DOI 10.3389/fnins.2014.00167.

Thomas C, Ye FQ, Irfanoglu MO, Modi P, Saleem KS, Leopold DA, Pierpaoli C. 2014. Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. *Proceedings of the National Academy of Sciences of the United States of America* **111(46)**:16574–16579 DOI 10.1073/pnas.1405672111.

Thomas Yeo BT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zöllei L, Polimeni JR, Fischl B, Liu H, Buckner RL. 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology* **106(3)**:1125–1165 DOI 10.1152/jn.00338.2011.

Tomassini V, Jbabdi S, Klein JC, Behrens TEJ, Pozzilli C, Matthews PM, Rushworth MFS, Johansen-Berg H. 2007. Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations. *The Journal of Neuroscience* **27(38)**:10259–10269 DOI 10.1523/JNEUROSCI.2144-07.2007.

Van den Heuvel M, Mandl R, Hulshoff Pol H. 2008. Normalized cut group clustering of resting-state FMRI data. *PLoS ONE* **3(4)**:e2001 DOI 10.1371/journal.pone.0002001.

Van den Heuvel MP, Sporns O. 2013. Network hubs in the human brain. *Trends in Cognitive Sciences* **17(12)**:683–696 DOI 10.1016/j.tics.2013.09.012.

Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, WU-Minn HCP Consortium. 2013. The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**:62–79 DOI 10.1016/j.neuroimage.2013.05.041.

Wakana S, Jiang H, Nagae-Poetscher LM, Van Zijl PCM, Mori S. 2004. Fiber tract-based atlas of human white matter anatomy. *Radiology* **230(1)**:77–87 DOI 10.1148/radiol.2301021640.

Wang L, Mruczek REB, Arcaro MJ, Kastner S. 2014. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex* Epub ahead of print DOI 10.1093/cercor/bhu277.

Ward JH. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58(301)**:236–244 DOI 10.1080/01621459.1963.10500845.

Wiggins JL, Peltier SJ, Ashinoff S, Weng SJ, Carrasco M, Welsh RC, Lord C, Monk CS. 2011. Using a self-organizing map algorithm to detect age-related changes in functional connectivity during rest in autism spectrum disorders. *Brain Research* **1380**:187–197 DOI 10.1016/j.brainres.2010.10.102.

Wig GS, Laumann TO, Petersen SE. 2014. An approach for parcellating human cortical areas using resting-state correlations. *Neuroimage* **93(Pt 2)**:276–291 DOI 10.1016/j.neuroimage.2013.07.035.

Wolf HC. 2000. Intranational home bias in trade. *Review of Economics and Statistics* **82(November)**:555–563 DOI 10.1162/003465300559046.

Xu R, Zhen Z, Liu J. 2010. Mapping informative clusters in a hierarchical [corrected] framework of FMRI multivariate analysis. *PLoS ONE* **5(11)**:e15065 DOI 10.1371/journal.pone.0015065.